Rules and Judgments in Statistics: Three Examples

Allan Stewart-Oaten

Ecology, Vol. 76, No. 6 (Sep., 1995), 2001-2009.

Stable URL:
http://links.jstor.org/sici?sici=0012-9658%28199509%2976%3A6%3C2001%3ARAJIST%3E2.0.CO%3B2-P

Ecology is currently published by The Ecological Society of America.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/esa.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Thu Aug 26 15:02:30 2004

Ecology, 76(6), 1995, pp. 2001–2009
© 1995 by the Ecological Society of America

RULES AND JUDGMENTS IN STATISTICS: THREE EXAMPLES[1]

ALLAN STEWART-OATEN
Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

Abstract. Statistical analyses are based on a mixture of mathematical theorems and judgments based on subject matter knowledge, intuition, and the goals of the investigator. Review articles and textbooks, aiming for brevity and simplicity, sometimes blur the difference between mathematics and judgment. A folklore can develop, where judgments based on opinions become laws of what "should" be done. This can intimidate authors and readers, waste their time, and sometimes lead to analyses that obscure the information in the data rather than clarify it. Three familiar examples are discussed: the choice between Normal-based and non-parametric methods, the use of multiple-comparison procedures, and the choice of sums of squares for main effects in unbalanced ANOVA. In each case, commonly obeyed rules are shown to be judgments with which it is reasonable to disagree. A greater stress on model selection, aided by informal methods, such as plots, and by informal use of formal methods, such as tests, is advocated.

Key words: analysis of variance; efficiency; multiple comparisons; nonparametric methods; sums of squares; validity.

INTRODUCTION

A common problem in statistical analyses in ecology is that judgments expressed in statistical papers, textbooks, and reviews are sometimes interpreted as mathematically rigorous, mandated rules, rather than as opinions more relevant to some cases than others. As a result, authors, reviewers, or editors sometimes select or require statistical analyses that, after much effort, obscure more than they clarify.

Reviews of statistical methods in biological journals have a valuable but difficult role. They may be the main route from theory to widespread application, so essential to the progress of both biology and statistics. The difficulties arise from the need to present the methods concisely and, as far as possible, painlessly. Mathematical derivations rarely appear, so the models and approximations on which the methods are based, and their limitations, may be downplayed. This can lead to a backlash later, when a method is seen to be exact only in special circumstances, such as an underlying Normal distribution, and is incorrectly tarred as "invalid" in all others. In addition, assertions tend to rely on proof by authority and to take the form of exhortations and instructions rather than statements of results. The distinction between fact and judgment is blurred. In time, some judgments become accepted as laws. These can be hard to challenge, since refutations may be viewed by statistical journals as "obvious" and by biological journals as "wrong."

I discuss three such judgments in areas recently reviewed in *Ecology* and *Ecological Monographs*: the choice of nonparametric vs. Normal-based procedures (Potvin and Roff 1993), the use of multiple testing methods (Day and Quinn 1989), and the choice of numerator sums of squares in unbalanced ANOVA (Shaw and Mitchell-Olds 1993). In each case, I offer and defend a judgment that disagrees in part with common beliefs described in these reviews.

NONPARAMETRIC VS. NORMAL-BASED PROCEDURES

"Nonparametric" or "distribution-free" methods arose mainly in the 1930s and 1940s from concerns about the validity of methods based on the Normal distribution, although Arbuthnot (1710) used the sign test to prove the existence of God. Two judgments commonly used to choose between nonparametric and Normal-based methods are: (1) Normal-based methods are valid only for Normal distributions, while nonparametric methods are valid for all distributions. (By "valid," I mean that the true probabilities of a test rejecting a true null hypothesis, or of a confidence interval covering an unknown parameter, equal the nominal alpha level or confidence.) One should test whether the data are Normal and, if they fail this test, use nonparametric methods. (2) Those nonparametric methods that are based on ranks (to make computations manageable) discard information, so they are less efficient than methods using the raw data. (By "efficient," I mean that the tests are less powerful, the estimates have

[1] Manuscript received 9 May 1994; revised 7 December 1994; accepted 27 January 1995.

greater standard errors, and the confidence intervals are wider, for a given sample size.)

Both beliefs are (in my "judgment") more wrong than right.

Potvin and Roff (1993) appear to support belief (1): "The main advantage of nonparametric methods is the absence of assumptions regarding the distribution underlying the observations." One way to check belief (1) is to compare results from the t table to those from the "randomization" distribution: the collection of all possible t values that can be obtained by keeping the *values* in the data but rearranging the *labels* ("treatment" or "control") attached to them. Under the null hypothesis of no effect, the value obtained from each unit was "pre-ordained," regardless of whether it was "treatment" or "control;" the chance calculated in a P-value (the significance level of the data) derives from the experimenter's deliberate randomized assignment of units to "treatment" or "control." Possibly the first such test was used by Fisher in 1935 (Fisher 1960) to criticize belief (1): for 15 data pairs from an experiment by Darwin (1876), the P value from the t table was almost identical to that obtained from the randomization distribution (i.e., all 2^{15} possible t values obtainable by labelling one member of each data pair "treatment" and the other "control"). This example does not constitute a proof, but Hoeffding (1952) proved that the standard and randomization two-sample t tests have the same validity and power for large samples.

A second check is to study the behavior of Normal-based methods when samples are drawn from non-Normal distributions. Numerous studies based on geometry (Efron 1969), asymptotics (e.g., Cressie and Whitford 1986), and simulations (e.g., Posten 1978, 1979), have indicated the broad validity of Normal-based methods for moderate non-Normal samples, even as low as $n = 5$. The exceptions are one-tailed tests on a single skewed distribution or on two skewed distributions with different skewnesses, variances, or sample sizes.

The advice to test for Normality before using Normal-based methods is almost paradoxical. If: "Normality is a myth: there never has been, and never will be, a Normal distribution" (Geary 1947), then the null hypothesis of Normality is always false. Whether it will be rejected in a given case will depend on the power of the Normality test, if a fixed cutoff (e.g., 0.05) is used. Obviously this power will be greater for large than for small samples. Thus pretesters will tend to accept Normality and use Normal-based methods when samples are small, but reject Normality and use other methods when samples are large. This is almost the opposite of what they "should" do since, when the underlying distribution is non-Normal, many Normal-based methods have high validity with large samples (courtesy of the Central Limit Theorem) but lower validity with small samples.

Belief (1) is wrong not only about the broad validity of Normal-based methods, but also about the universal validity of nonparametric methods. Unless there is deliberate randomized assignment of experimental units to treatments (as in the Darwin example), the latter are almost always based on the assumption that, under the null hypothesis, all observations are independent draws from the same distribution. When used for confidence intervals, e.g., for the difference between the means of two populations, they usually assume that the populations are identical except for a location or shift parameter. These assumptions of identical distributions require equal variances under different treatments. Hoeffding's (1952) results can be extended to show that the randomization and standard (equal variances) t tests have the same validity and power when variances are unequal (Romano 1990): both are asymptotically *in*valid if the sample sizes are unequal (i.e., if their ratio does not converge to unity). Fligner and Policello (1981) show that the Wilcoxon-Mann-Whitney test can be invalid when comparing distributions having different variances, and suggest an adjustment similar to the Welch-Satterthwaite modification of the t test. Without this adjustment, the Wilcoxon is likely to be *less* valid than the modified t: the "rule" that nonparametric methods "should" be used when variances are unequal is especially unfortunate.

The t and the Wilcoxon two-sample tests are actually testing different things: $E(X) = E(Y)$ for the t, but $P(X > Y) = 0.5$ for the Wilcoxon. Each implies the other if the distributions are symmetric, or have the same shape and spread. But if the two distributions are differently skewed (as in the example given in Table 2 and text of Potvin and Roff [1993]), or have the same skew but different variances, the tests are hard to compare: one hypothesis could be true while the other is false (or both could be false, but in opposite directions!), and neither test is strictly valid.

Overall, "(1) *unless* randomization has been performed, the 'distribution free' tests do not possess the properties claimed for them, and (2) *if* randomization is performed, standard parametric tests such as the t test usually supply adequate approximations" (Box et al. 1978: 104).

Potvin and Roff's (1993) review provides a good survey of evidence contradicting belief (2). It was first attacked by Pitman (1948), who formalized ideas of the efficiency of tests, and later in classic papers by Hodges and Lehmann (1956) and Chernoff and Savage (1958). In brief, nonparametric methods are often only slightly less efficient than Normal-based methods when the true underlying distribution is Normal, but can be much more efficient when it is not (Lehmann 1975).

Ironically, this is a by-product of the reduction of information to ranks, a device aimed originally not at increasing efficiency but at making computations manageable in pre-computer environments. The use of ranks reduces the effect of extreme observations, which carry less information than moderate observations for non-Normal distributions, and can be seriously misleading. Thus the greater efficiency derives not from being distribution-free (i.e., using randomization distributions rather than theoretical sampling distributions), but from the use of statistics chosen for computational convenience.

But extreme observations can also be downweighted in parametric approaches. Trimmed means, modified maximum likelihood (Tiku et al. 1986), and other "*L*," "*R*," or "*M*" estimators (Andrews et al. 1972, Huber 1981) all have only slightly larger variances than the sample mean when the underlying distribution is Normal, but often much smaller variances when it is not. The smaller variances lead to smaller confidence intervals and more powerful tests (Gross 1976, Kafadar 1982). These "robust" methods often produce estimates of means, regression slopes, etc., more easily than rank-based methods. Messy computations may be needed to derive confidence intervals and tests from them (e.g., to estimate variances), but these can be automated and the results appear to be valid for non-Normal distributions. Their main drawbacks may be (1) an author may be suspected (perhaps rightly) of selecting the method whose results best fit his pet theory, and (2) like rank methods, they estimate a variety of different things when distributions are skewed.

In summary, the main advantage of nonparametric methods is not greater validity, which is usually slight, but greater efficiency, which is often considerable, although it depends on the underlying distribution. Rote resort to nonparametric methods because of suspected non-Normality, or the failure of a pre-test, is not a good strategy, especially if it leads to testing a plainly implausible null hypothesis rather than estimating a meaningful parameter. Rather than pre-test for Normality, it makes sense to plot the data. Severe kurtosis might suggest a nonparametric or a robust parametric procedure, while severe skewness might suggest a transformation, a topic too large to discuss here. If the "identical distributions" assumption can be trusted, then nonparametric methods seem especially useful when samples are small (so computations are simple and Normal-based approximations may be seriously inaccurate), their target parameters are of particular interest (e.g., the median), a complicated statistic is needed (e.g., to estimate a mode), or there is a need for both efficiency and simplicity (e.g., high kurtosis and an audience likely to be suspicious of exotic methods).

Multiple Comparisons

Most studies involve several statistical tests or confidence intervals. The chance that at least one of these makes an error (a false rejection or an interval that fails to cover the true value), will be larger than the chance of an error on any particular one. Also, some "unplanned" tests or intervals may have been constructed only because the investigator noticed something odd about the data: in effect, many tests were carried out mentally, but only the most "significant" was reported. Multiple-comparison methods allow simultaneous inference with a prespecified overall error rate (probability of at least one false rejection or incorrect interval) in cases like these.

Several different judgments have been made about the choice between overall rates controlled by multiple-comparison methods and the usual comparison-wise rates that consider each test or interval in isolation. Scheffé (1959: 66, 80) recommends using overall rates for all inferences on a single data set, following a significant ANOVA *F* test. Snedecor and Cochran (1989: 234) recommend overall rates for unplanned tests or intervals, but comparison-wise rates for those that were planned before the data became available. Sokal and Rohlf (1981: 233, 241) and Day and Quinn (1989: 449) recommend comparison-wise rates for a set of planned orthogonal comparisons, but overall rates for other planned and all unplanned comparisons. (In one-way ANOVA, a contrast of the means is a linear function, $\Sigma_j c_j \mu_j$, where the c_js are constants with $\Sigma_j c_j = 0$; two contrasts, using $\{c_{1j}\}$ and $\{c_{2j}\}$, are orthogonal if $\Sigma_j c_{1j} c_{2j}/r_j = 0$, where r_j is the number of observations on treatment *j*). Finally there is a viewpoint rather rare in biology but common among statisticians:

"Multiple comparison methods have no place at all in the interpretation of data" (Nelder 1971).

"I recommend therefore that multiple comparison methods be avoided; that the idea of experiment-wise error rates be retained, but only as a general principle" (Mead 1988).

My judgment is closest to these last. A list of reasons for avoiding multiple comparisons may be useful, provided it is kept in mind that these reasons are *judgments* rather than theorems, not all of them apply in all cases, not all opponents of multiple comparisons have the same list, and there may be cases where none of these reasons apply and multiple comparisons are useful. Miller (1981) and the review by Day and Quinn (1989) are good guides in such cases and exceptions to Finney (1990), who suggests rejecting all multiple-comparisons papers as "rarely more useful than a horoscope."

The role of significance tests

It is an illusion to see testing as a system of objective, automated decision-making: "an effect is real if and only if the null hypothesis is rejected at the 0.05 level." Multiple-comparison methods are in part an attempt to maintain this illusion by protecting the integrity of the test level. The task is impossible. Dozens of scientists are working every day on effects that, unknown to them, are small, unimportant, or non-existent. By chance, a few will get "significant" results. These are far more likely to be submitted and published than the others. Unpublished dissertations can correct the balance a little, but the integrity of 0.05 is hopelessly lost.

In any case, few of us would make up our minds on the basis of a single test on a single data set. We will want to consider evidence on both presence and size from several sources, and also our biological intuition and knowledge of mechanisms, similar systems and species, etc. In selecting a best treatment (e.g., in medicine), we will want to assess costs, availability, ease of use, side effects, etc. (Anscombe 1985).

Probably no single number can combine all this information (though Bayesians and some meta-analysts may disagree). An accept/reject result at the 0.05 level is only one of several summaries to consider, perhaps one of the least useful. It contains less information than a level of significance (*P*) or a confidence interval. Plots, averages and other estimates, mean squares, standard deviations and *F* values, and well-organized tables may also tell us more about a particular data set. Information from other data sets, or about mechanisms from quite different studies, may tell us more still. Thus, even with an uncontaminated test level, multiple-comparison tests devote great effort to delivering one of the least important data summaries, and inflating its importance.

Interpretability

The comparison-wise error rate is the probability of rejecting a true null hypothesis. The experiment-wise rate is the probability of rejecting any of a set of true null hypotheses tested in an experiment. We could also define error rates for papers, studies, or (for journal editors) issues or volumes. Recent issues of the Royal Statistical Society's *RSS News* have (as a joke) suggested controlling lifetime error rates: the first test of your career is done at level 0.05/2, the next at 0.05/4, and so on. None of these rates provides perfect protection; e.g., none protects a reader against the selection bias mentioned above. Once test results lose their sacred aura, and are recognized as only one, perhaps minor, summary of a single data set, it makes sense to choose on the basis of simplicity and interpretability. This points to the comparison-wise rate. An additional consistency argument (Saville 1990) is that, for all other rates, the test result depends not only on the data relevant to the question, but also on irrelevant information such as the number of other questions studied with it, and perhaps their results.

Which experiment-wise rate?

Many multiple-comparison procedures are used only if an ANOVA *F* test first rejects the hypothesis that all contrasts in the class are zero (e.g., that all means are equal). Thus the long-run fraction of errors (tests that falsely reject, intervals that fail to cover the true parameter) will be the *conditional* probability of a wrong answer *given* the significant *F* test, not the usual unconditional value. This conditional probability depends on unknown parameters, but is always $\geq\alpha$, sometimes much greater, for Scheffé (1959) tests and confidence intervals (Olshen 1973).

Difficulty

Multiple-comparison tests are relatively difficult to understand and carry out. This has three bad effects: they may distract investigators from more effective ways of assessing their data, they may be used inappropriately, and unsophisticated readers attribute more importance to them than they deserve. Mead (1988) gives startling examples of published work in which a clutter of incomprehensible multiple comparisons served to obscure biologically significant patterns that plotting made immediately apparent. The most frequently cited statistical paper for 1945-1988, ranked 24th in all of science (Garfield 1990), was Duncan (1955), whose multiple-range test is almost always unsuitable and inappropriately applied because it does not control experiment-wise error rates, or any other error rate that can be succinctly described (Day and Quinn 1989).

Rigidity

The choice of test level is usually arbitrary: it is not derived mathematically from generally agreed criteria. It makes sense to use a small α if Type 1 errors are very serious and Type 2 errors only mildly so, and to use a larger α in the reverse case. Since seriousness often depends on the user, levels of significance (*P*) are preferable to accept/reject decisions. They are also more informative as partial data summaries. Multiple-comparison tests usually ignore this: all tests are treated equally.

Planned comparison-wise *P* values can be approximately converted to experiment-wise *P* values, as an informal caution against overinterpretation. The probability of getting a comparison-wise *P* value of 0.03 in 10 tests of true null hypotheses is $\leq 10(0.03) = 0.3$, using the Bonferroni inequality (and often close to this value). But if a given hypothesis is not rejected at $\alpha = 0.05$ by a multiple-comparison test, we can do little more than say that its comparison-wise *P* value is

$\geq$0.005, possibly very much greater. This point has less force for unplanned comparisons that, in theory, are selected from infinitely many possible comparisons. But this theory may be unrealistic. Usually, only a few pairwise differences, and a few differences between averages of one type and averages of another type, have reasonable biological interpretations. Thus, even here, it may be possible to allow approximately for the number of tests that could have been made.

Power

For a given α, a given null hypothesis is less likely to be rejected by a multiple-comparison test than by a single test. This reduces Type I error but increases Type II error. There is no obvious reason for this to be a good trade-off. It is clearly bad if, as a result, "significance" requires a virtually (or literally) impossible effect size: the test result will contain no information. It is even less justified if a general F test has already discredited the presumption that all null hypotheses are true. A related problem is that the extreme test statistics required for "significance" will be in the distant tails of the null distribution, where approximations like the Central Limit Theorem may not apply for moderate samples.

The experimenter's intentions as a datum

A contrast is to be tested by one method if it was planned but by another if it was not. Methods for planned tests depend on the number of tests planned: if k tests were planned, the Bonferroni procedure tests at level α/k. How is a reader to know whether an author is being honest about his intentions? Also, why should the reader care? Perhaps the author did not think of a particular contrast until he saw the data, so felt compelled to use the Scheffé method. But the reader may have thought of it immediately because of other observations she had made. As a result, she may have been interested in this comparison and no others. Why should his prior beliefs take precedence over hers?

Orthogonality

Some authors propose special treatment for orthogonal contrasts because their estimates, $\Sigma\ c_{1j}y_{j\cdot}$ and $\Sigma\ c_{2j}y_{j\cdot}$ (the dots indicate averaging over the missing subscript), are independent. However, independence requires equal variances unless $c_{1j}c_{2j} = 0$ for all j. Also, the orthogonality condition only ensures zero correlation; this implies independence if the errors are Normal, but not otherwise. Even then, *inferences* are not usually independent: they use the same variance estimate, the residual mean square. And even if the tests were independent, the chance of rejecting at least one true null hypothesis increases with the number of tests just as inexorably as for correlated tests.

An emphasis on orthogonal contrasts risks allowing the needs of the statistical analysis to determine the scientific questions to be studied: the tail wags the dog. Snedecor and Cochran (1989) and Day and Quinn (1989) contend that orthogonal contrasts provide separate answers to separate biological questions, but this implies that the questions "Is $\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$?" and "Is $\mu_1 - \mu_2 + \mu_3 - \mu_4 = 0$?" are separate biological questions if variances and samples sizes are equal, but not otherwise.

Orthogonality is important, but the time to consider it is not when the data are analyzed but when the experiment is being designed, so that estimates of the contrasts of main interest will (as far as possible) be uncorrelated and have small variances.

In summary, it would be risky to claim that multiple-comparison methods should never be used: one can't think of everything. But they seem best suited for the privacy of one's own lab, where (along with some biological thought) a small set of methods (e.g., Bonferroni, Scheffé, and Tukey–Kramer; see Day and Quinn 1989) might reduce overinterpretation without great effort. For publication, comparison-wise methods seem preferable, with a warning that the use of many inferences raises the overall probability of misleading results. In some cases, e.g., fishing expeditions where the tests are used mainly as an exploratory tool, the warning could be reinforced by a few multiple-comparison results, to indicate roughly the amount of allowance needed.

Sums of Squares for ANOVA Tests

With what F ratios should effects be tested in an unbalanced k-way ANOVA? This is a perennial and potentially frustrating problem, as those who have struggled with Types I, II, and III sums of squares in SAS will testify. Several authors, and the SAS manuals, appear to favor Type III sums of squares. In explaining my disagreement, I focus mainly on the two-way setup. Its Full model, without restrictions, can be written in either of two standard forms. One is the "cell means" model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} is the kth observation ($k = 1, 2, \ldots, n_{ij}$), using row treatment i (e.g., fertilizer i: $i = 1, 2, \ldots, f$) and column treatment j (e.g., variety j: $j = 1, 2, \ldots, v$), μ_{ij} is the mean for this combination of treatments, and ϵ_{ijk} is "error" due to other sources of variation.

The other form is the "main effects and interactions" ("effects") model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{2}$$

where μ is the "grand mean," α_i and β_j are the main effects of the ith fertilizer and of the jth variety, and

TABLE 1. Standard tests of H_0: all $\alpha_i = 0$ (e.g., no fertilizer effects).*

SAS Type	Complete model	Cell means hypothesis tested
Type I	(μ, A)	H_{10}: $\mu_{i\#}$ is the same for all i
Type II	(μ, A, B)	H_{20}: $\mu_{i\#} = \Sigma_j n_{ij}\mu_{\#j}/n_{i+}$ for all i
Type III	(μ, A, B, AB)	H_{30}: $\mu_{i\cdot}$ is the same for all i

* $\mu_{i\cdot}$ given by Eq. 3; $\mu_{i\#}$ given by Eq. 5. Type I ss assumes the model is entered in SAS as "*A B A***B*."

$(\alpha\beta)_{ij}$ is their interaction. This model is over-parameterized, but one can reasonably define $\mu = \mu_{\cdot\cdot}$, $\alpha_i = \mu_{i\cdot} - \mu$, $\beta_j = \mu_{\cdot j} - \mu$ and $(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$, where the dots indicate unweighted averages over the missing subscript(s). E.g.,

$$\mu_{i\cdot} = \Sigma_j \mu_{ij}/v, \qquad (3)$$

so α_i is (average of the means using fertilizer i) − (average of all means). These definitions introduce the "usual" side conditions, $\Sigma\alpha_i = \Sigma\beta_j = \Sigma_i(\alpha\beta)_{ij} = \Sigma_j(\alpha\beta)_{ij} = 0$, so there are really only $f - 1$ α's, $v - 1$ β's and $(f - 1)(v - 1)$ $(\alpha\beta)$'s. (These effects parameters can be defined in other ways, leading to different side conditions.)

The usual null hypotheses posit sets of linear relations among the means, μ_{ij}, of Eq. 1. In principle, any relation can be tested, but the only ones tested in practice are more easily expressed in terms of the effects model, Eq. 2, as "all α's (or β's or $(\alpha\beta)$'s) are zero." We can describe the corresponding models by listing the components, using μ, A, B, and AB to indicate the inclusion of μ, the α's, the β's, and the $(\alpha\beta)$'s. Thus the model $Y_{ijk} = \mu + \alpha_i + (\alpha\beta)_{ij} + \epsilon_{ijk}$, which assumes that all β_j's are zero, is designated (μ, A, AB).

Assuming (in decreasing order of importance) independent observations, equal variances, and Normality, the hypotheses that there are no interactions, no row treatment effects, or no column treatment effects, are tested by comparing the fit of any "Complete" model known (or believed) to be true to that of the same model with the parameters under test omitted. There are eight possible Complete models for testing "H_0: all α_i's are zero:" (μ, A, B, AB), (μ, A, B), (μ, A), (μ, A, AB), (A, B, AB), (A, B), (A, AB) and (A); in each case the Reduced model is obtained by omitting "A." The Complete model is sure to fit better, but the improvement may be due only to chance. To test this, the difference in fits is compared to an independent estimate of σ^2, the variance of the ϵ_{ijk}'s, obtained by comparing the sum of squares of the observations with the sum of squares of the fitted values for any "Trusted" model believed to be true. The F test statistic is:

$$F = \{(\text{Complete ss} - \text{Reduced ss})/\text{dfn}\}/\{(\text{Observed ss} - \text{Trusted ss})/\text{dfd}\} \qquad (4)$$

where Complete ss = Sum of squares of fitted values using the Complete model, etc., and dfn (= the number of parameters under test) and dfd (= n − the number of parameters in the Trusted model) are the degrees of freedom of the numerator and denominator.

The problem is, what should be the Complete and Trusted models? We focus on the Complete model, which is usually the main concern. The Trusted model is often the Complete model, but can be larger (contain more parameters). E.g., it could be the Full model, containing all effects and interactions ([μ, A, B, AB] in the two-way case). We also focus on tests of main effects in the two-way model: there is little dispute that (μ, A, B, AB) is the appropriate Complete model for tests of interactions. To avoid extraneous complications, we assume all $n_{ij} > 0$.

Shaw and Mitchell-Olds (1993) advocate use of (μ, A, B, AB) for testing main effects too. This follows the advice of Speed et al. (1978), who argue that other choices test the wrong hypotheses. They describe the hypotheses not in terms of the α's or β's of the effects model, Eq. 2, but in terms of the μ_{ij}'s of the cell means model, Eq. 1. For a given choice of Complete model, the hypothesis tested is the relation among the μ_{ij}'s for which, if there were no errors, Complete ss = Reduced ss. These relations can be expressed in terms of row and column averages of the μ_{ij}'s ($\mu_{i\cdot}$ and $\mu_{\cdot j}$), and of row and column *weighted* averages, the weights being the cell counts, n_{ij}:

$$\mu_{i\#} = \Sigma_j\, n_{ij}\mu_{ij}/n_{i+} \text{ and } \mu_{\#j} = \Sigma_i\, n_{ij}\mu_{ij}/n_{+j}, \qquad (5)$$

where $n_{i+} = \Sigma_j\, n_{ij}$ and $n_{+j} = \Sigma_i\, n_{ij}$.

Of the eight possible Complete models for testing "H_0: all α's are zero," given in the paragraph preceding Eq. 4, only the first three are used in practice: it is almost never realistic to suppose that $\mu = 0$, and the model (μ, A, AB) is rarely plausible. Thus, Table 1 gives the standard choices.

Speed et al. (1978) argue that H_{30} (and its counterpart for the β's) "seem to be reasonable," while H_{10} and especially H_{20} are "not easy to understand" and "very difficult to justify." This is clearly a set of judgments about what "seems reasonable," rather than a set of mathematical results. But there are reasonable counterarguments. When Hocking (1982) made similar arguments in connection with the Analysis of Covariance, Cox and McCullagh (1982) replied that he "appears to favor estimating and testing main effects in the presence of interaction, a thing we consider rarely physically meaningful.''

If there are no AB interactions, H_{20} is equivalent to H_{30}, and both to ``all α_i's are equal.'' For example, $\mu_{i\#}$ becomes $\mu + \alpha_i + \Sigma_j n_{ij}\beta_j/n_{i+}$ while $\Sigma_j n_{ij}\mu_{\#j}/n_{i+}$ becomes $\mu + \Sigma_j\Sigma_k n_{kj}n_{ij}\alpha_k/n_{+j}n_{i+} + \Sigma_j n_{ij}\beta_j/n_{i+}$, so $\alpha_i = \Sigma_j\Sigma_k n_{kj}n_{ij}\alpha_k/n_{+j}n_{i+}$, a weighted average with no zero weights; this holds for all i so all α_i's must be equal: e.g., it would be impossible for the ``largest'' $\alpha$ to be a weighted average of the others. Thus $H_{20}$ is ``not easy to understand'' if there are interactions, but seems easy to justify otherwise, since it becomes ``all $\alpha$'s are 0'' or (in the means model) ``$\mu_{ij} = \mu._j$ for all i and j.'' Similarly, if there are neither AB nor B effects, then H_{10} is equivalent to the other two, and ``seems'' quite reasonable.

Thus, H_{30} is clearly preferable only when there are interactions. But all three H_0's ``seem'' likely to be uninteresting then. Shaw and Mitchell-Olds (1993) give an example in which Y = tree height at the end of a study, A refers to removal or non-removal of conspecifics within a certain distance, and B is initial height (Small or Large). With interactions, H_{30} would mean that removal increases height of one B class (say, the initially small trees) but decreases that of the other, the two effects being exactly equal. This seems an impractical result, more an unlikely coincidence than anything else: the averages are unlikely to remain equal if we choose a different dividing line between Small and Large initial heights, or divide initial heights into three classes rather than two. In experiments with less arbitrary categories, adding a new category (e.g., a new seed type) to the Bs would also usually make the means of the As unequal. Once we know the removal effect varies, the main effects are usually of little interest compared to the mechanisms or consequences. This is the point made by Cox and McCullagh (1982), but obscured by the cell means model: the Type III SS is ``obviously'' best for a test of main effects only when it makes little sense to test main effects at all.

If we test main effects only after deciding that there are no interactions, then valid F tests of ``all $\alpha$'s are equal'' (or ``all $\mu_i.$'s are equal'') are obtained with either (μ, A, B) or (μ, A, B, AB) as the Complete model. It then seems reasonable to choose on the basis of power. This leads to choosing (μ, A, B), as Shaw and Mitchell-Olds (1993) remark. If there are neither interactions nor B effects, then the Type I SS, using (μ, A), is more powerful than either. In both cases, the power argument also leads to choosing the Complete model as the Trusted model in Eq. 4, in order to maximize dfd.

This suggests that the best procedure for determining the true model in a two-way unbalanced ANOVA is usually (1) test for interactions, using the Type III SS; (2) if there are no interactions, check A or B effects, whichever seems a priori more likely to be absent, using the Type II SS; and (3) if these main effects seem to be absent, check the remaining main effects using the Type I SS, otherwise use the Type II SS. If interactions are present, main effects would usually not be tested. The emphasis is on models: we begin with a large one and simplify it in steps, each time testing whether the data justify removing, from the model currently accepted, the complicating factor judged least likely to be present. A more general scheme, involving a ``baseline'' model that can be updated iteratively, is suggested by Cox (1984) for higher-way layouts. The SSS in this scheme would often be different from all of the ``Types'' routinely offered by packages like SAS.

There may be exceptional cases where Type III SSS make sense. A set of treatments might sometimes be expected to have exactly opposite effects on males and females, or on the left and right sides. I know of no example, but the claim that some marine structures ``increase'' density merely by attracting fish from elsewhere might lead to one. If interactions, though ``significant,'' seem small, it may make sense to see whether some treatments (e.g., fertilizers) do consistently better than others, though Type III tests for fertilizer effects are not necessarily the best way to do this. If treatment 1 of both the A and B groups is a control, it might make sense to define effects in terms of differences from it: as $\alpha_i = \mu_{i1} - \mu_{11}$, $\beta_j = \mu_{1j} - \mu_{11}$, and $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11}$; if it is plausible that A treatments have no effects except when combined with B treatments, then models like (μ, AB) may be reasonable and SSS different from any of the ``Types'' called for. These examples suggest that generalizations about appropriate SSS may be less useful than the injunction to ``think about the model.''

It should be stressed that ``using'' a SS does not mean decisions ``should'' be based only on formal tests. Plots, conformity with other data and information, plausibility of mechanisms, and apparent sizes of effects are also relevant, and ``should'' perhaps play a larger role than tests. Cox (1984) notes that his procedure is ``close to a severely constrained form of stepwise regression;'' using these other factors would bring it closer to a constrained form of variable selection in regression, where plots, measures like C_P (Mallows 1973), and subject matter information are all brought to bear. Recent work on exploratory approaches (Hoaglin et al. 1991) is relevant here.

Discussion

The aim of this paper is not to develop new orthodoxies of the universal validity of Normal-based procedures, universal avoidance of multiple-comparison methods or the appropriate sums of squares in unbalanced ANOVA, nor to suggest there are no rules of inference at all. It is to suggest that the practice of statistics in ecology is sometimes too rigid. "Sensible" opinions are treated as mandatory rules, frustrating authors, who may be required to use methods they think inappropriate, and confusing readers.

One possible reason for this is an aura of exactitude and universality inherited from mathematics, although almost all inference is approximate not only because of the use of limiting distributions but also because of model uncertainty, treatment–subject non-additivity, and sampling that is random in at best a limited slice of time and space. Another is the narrow options and focus of most canned packages: like many students they "have no idea that the summary of an experiment is not the anova table but tables of means and standard errors" (Nelder 1994). (Another example: neither SAS nor SYSTAT gives a confidence interval for the difference of two means when the variances are unequal.) A third is oversimplification, avoidance of models, narrowness of focus, and proof by authority in many reviews and textbooks.

Confusion of Popperian tests of a theory with "Fisherian" tests of a null hypothesis is a fourth reason, though this unfairly oversimplifies Fisher's role. The former subjects theories to grave danger of refutation, while the latter too often merely supports them by rejecting a null hypothesis that no one believed in the first place. "Since the null hypothesis is quasi-always false, tables summarizing research in terms of patterns of 'significant differences' are little more than complex, causally uninterpretable outcomes of statistical power functions" (Meehl 1978). Meehl's comment that "the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" applies to other disciplines too.

Statistical techniques are intended to clarify: to separate signal from noise, reveal patterns, and tease out systematic or consistent relationships and tendencies that might otherwise be hidden in individual variation, measurement error, and random accidents. Some hypothesis testing can help, since some null hypotheses could be true: some birds may make no distinction at all between their own eggs and those of a parasite (Rothstein 1986), and interactions may sometimes all be exactly zero. Goodness-of-fit tests often do subject well-defined models to danger of refutation, rather than confirming vague ones by rejecting something implausible. But much of statistics is not formal inference but an often-iterative pattern of design, summary, description, and display, and linking results of the present study to past studies, known mechanisms, and biological intuition. Most of formal inference is not hypothesis testing but model construction, selection, and checking (formal and informal), estimation of parameters and standard errors, or calculation of confidence regions or of Bayesian posterior distributions of parameters (Box 1980).

When "statistics" gets reduced to "statistical inference" and then to hypothesis testing (Freedman et al. 1991 is an outstanding exception), it can become its opposite: not a way to reveal and clarify but to obscure and terrify. Statistical consultants are then asked not to help discover truth, but to produce mumbo jumbo, of no interest to the client and distracting to the audience, to pacify a referee.

A program to promote better statistical practices would not only be a large undertaking, but might become just as fossilized as present practice. Some suggestions might be useful: more use of informal, non-inferential techniques (plots, summary statistics, and tables, as described in several books on Exploratory Data Analysis); a much greater emphasis on model selection, justification, and checking when formal techniques are used; a healthy skepticism of the word "should," especially in non-mathematical reviews and texts; a reduction in significance testing of implausible null hypotheses; and a general aim to use statistical methods to simplify and reveal, as well as to measure uncertainty. It may be helpful to recognize that, although there are "wrong" analyses (incorrectly derived from explicit models, or based—often implicitly—on assumptions or models known not to be approximately true), there are usually several "right" ones, depending on the aims of the analysis and the knowledge, intuition, and uncertainties (i.e., the models) of the investigator. Except for some extreme Bayesians, we are rarely likely to have a complete and unique set of data analysis rules—or, one hopes, of regulations—but rather a set of sensible guidelines that, though supported by mathematical results, must be flexible enough to accommodate a variety of data-gathering setups and individual interests. "If inference is what we think it is, the only precept or theory which seems relevant is the following: 'Do the best you can.' This may be taxing for the old noodle, but even the authority of Aristotle is not an acceptable substitute" (LeCam 1977).

Acknowledgments

I thank Ed McCauley, Bill Murdoch, and Sue Swarbrick for helpful comments on two drafts, Steve Rothstein and Steve Gaines for useful further suggestions, and two anonymous referees for many others. This work was supported in part by the Minerals Management Service, U. S. Department of the Interior, under Minerals Management Service Agreement Number 14-35-0001-30471 (The Southern California Educational Initiative). The views and interpretation contained in this document are those of the author and should

not be interpreted as necessarily representing the official policies, either express or implied, of the U. S. Government.

Literature Cited

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. 1972. Robust estimates of location. Princeton University Press, Princeton, New Jersey, USA.

Anscombe, F. J. 1985. Review of "Simultaneous Statistical Inference" (Miller 1981). Journal of the American Statistical Association **80**:250.

Arbuthnot, J. 1710. An argument for divine providence, taken from the constant regularity observed in the births of both sexes. Philosophical Transactions **27**:186–190.

Box, G. E. P. 1980. Sampling and Bayes inference in scientific modeling and robustness (with discussion). Journal of the Royal Statistical Society **A143**:383–430.

Box, G. E. P, W. G. Hunter, and J. S. Hunter. 1978. Statistics for experimenters. John Wiley & Sons, New York, New York, USA.

Chernoff, H., and I. R. Savage. 1958. Asymptotic Normality and efficiency of certain nonparametric test statistics. Annals of Mathematical Statistics **29**:972–994.

Cox, D. R. 1984. Interaction. International Statistical Review **52**:1–31.

Cox, D. R., and P. McCullagh. 1982. Some aspects of analysis of covariance. (With discussion.) Biometrics **38**:541–561.

Cressie, N. A. C., and H. J. Whitford. 1986. How to use the two-sample *t* test. Biometrical Journal **28**:131–148.

Darwin, C. 1876. The effects of cross- and self-fertilization in the vegetable kingdom. John Murray, London, England.

Day, R. W., and G. P. Quinn. 1989. Comparisons of treatments after an analysis of variance in ecology. Ecological Monographs **59**:433–463.

Duncan, D. B. 1955. Multiple range and multiple *F* tests. Biometrics **11**:1–42.

Efron, B. 1969. Student's *t* test under symmetry conditions. Journal of the American Statistical Association **64**:1278–1302.

Finney, D. 1990. Letter to the editor. Biometrics Bulletin **7**:2.

Fisher, R. A. 1960. The design of experiments. Seventh edition. Oliver and Boyd, Edinburgh, Scotland.

Fligner, M. A., and G. E. Policello II. 1981. Robust rank procedures for the Behrens–Fisher problem. Journal of the American Statistical Association **76**:162–168.

Freedman, D., R. Pisani, R. Purves, and A. Adhikari. 1991. Statistics. Norton, New York, New York, USA.

Garfield, E. 1990. The most-cited papers of all time, Science Citation Index 1945-1988. Part 1B. Superstars new to the SCI top 100. Current Contents **8**:3–13.

Geary, R. C. 1947. Testing for Normality. Biometrika **34**: 209–242.

Gross, A. M. 1976. Confidence interval robustness with long-tailed distributions. Journal of the American Statistical Association **71**:409–416.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 1991. Fundamentals of exploratory analysis of variance. John Wiley & Sons, New York, New York, USA.

Hocking, R. R. 1982. Discussion of Cox and McCullagh (1982). Biometrics **38**:559–561.

Hodges, J. L., Jr., and E. L. Lehmann. 1956. The efficiency of some nonparametric competitors of the *t*-test. Annals of Mathematical Statistics **27**:324–335.

Hoeffding, W. 1952. The large sample power of tests based on the permutation of observations. Annals of Mathematical Statistics **23**:169–192.

Huber, P. 1981. Robust statistics. John Wiley & Sons, New York, New York, USA.

Kafadar, K. 1982. Using biweight m-estimates in the two-sample problem. Part 1: symmetric populations. Communications in Statistics, Theoretical Methods **11**:1883–1901.

LeCam, L. 1977. A note on metastatistics or 'an essay toward stating a problem in the doctrine of chances.' Synthese **36**: 133–160.

Lehmann, E. L. 1975. Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco, California, USA.

Mallows, C. L. 1973. Some comments on C_p. Technometrics **15**:661–675.

Mead, R. 1988. The design of experiments. Cambridge University Press, Cambridge, England.

Meehl, P. E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology **48**:806–834.

Miller, R. G., Jr. 1981. Simultaneous statistical inference. Springer-Verlag, New York, New York, USA.

Nelder, J. A. 1971. Contribution to the Discussion of O'Neill, R. T., and B. G. Wetherill. 1971. The present state of multiple comparison methods. Journal of the Royal Statistical Society, 'B', **33**:218–241.

Nelder, J. A. 1994. Science—a teaching framework. Royal Statistical Society News **21**:1–2.

Olshen, R. A. 1973. The conditional level of the *F* test. Journal of the American Statistical Association **68**:692–698.

Pitman, E. J. G. 1948. Lecture notes on nonparametric statistics. Columbia University, New York, New York, USA.

Posten, H. 1978. The robustness of the two-sample *t*-test over the Pearson system. Journal of Statistical Computation and Simulation **6**:295–311.

———. 1979. The robustness of the one-sample *t*-test over the Pearson system. Journal of Statistical Computation and Simulation **9**:133–149.

Potvin, C., and D. A. Roff. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology **74**:1617–1628.

Romano, J. P. 1990. On the behavior of randomization tests without a group invariance assumption. Journal of the American Statistical Association **85**:686–692.

Rothstein, S. I. 1986. A test of optimality: egg recognition in the eastern phoebe. Animal Behavior **34**:1109–1119.

Saville, D. J. 1990. Multiple comparison procedures: the practical solution. The American Statistician **44**:174–180.

Scheffé, H. 1959. The analysis of variance. John Wiley & Sons, New York, New York, USA.

Shaw, R. G., and T. Mitchell-Olds. 1993. ANOVA for unbalanced data: an overview. Ecology **74**:1638–1645.

Snedecor, G. W., and W. G. Cochran. 1989. Statistical methods. Eighth edition. Iowa State University, Ames, Iowa, USA.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. Second edition. Freeman, New York, New York, USA.

Speed, F. M, R. R. Hocking, and O. P. Hackney. 1978. Methods of analysis of linear models with unbalanced data. Journal of the American Statistical Association **73**:105–112.

Tiku, M. L., W. Y. Tan, and N. Balakrishnan. 1986. Robust inference. Marcel Dekker, New York, New York, USA.